Predicting Protein-Protein Interactions from Protein Sequences Using Phylogenetic Profiles

Ömer Nebil Yaveroğlu, *Member, IEEE,* and Tolga Can

Abstract—In this study, a high accuracy protein-protein interaction prediction method is developed. The importance of the proposed method is that it only uses sequence information of proteins while predicting interaction. The method extracts phylogenetic profiles of proteins by using their sequence information. Combining the phylogenetic profiles of two proteins by checking existence of homologs in different species and fitting this combined profile into a statistical model, it is possible to make predictions about the interaction status of two proteins.

For this purpose, we apply a collection of pattern recognition techniques on the dataset of combined phylogenetic profiles of protein pairs. Support Vector Machines, Feature Extraction using ReliefF, Naive Bayes Classification, K-Nearest Neighborhood Classification, Decision Trees, and Random Forest Classification are the methods we applied for finding the classification method that best predicts the interaction status of protein pairs. Random Forest Classification outperformed all other methods with a prediction accuracy of 76.93%.

Keywords—Protein Interaction Prediction, Phylogenetic Profile, SVM , ReliefF, Decision Trees, Random Forest Classification

I. Introduction

Identification of protein-protein interactions (PPIs) is important for understanding protein functions and biological processes in a cell. Representing the set of pairwise interactions as protein interaction networks is useful for understanding the cellular functions at a systems level. There are various methods used for predicting interaction of protein pairs. Experimental methods, computational inference methods, and protein interaction databases are the main data sources used for finding out interaction information of protein pairs. All these methods have a number of disadvantages. Experimental methods are rather expensive and they can find out a small number of interactions which are specifically targeted. Protein interaction databases are especially useful for generating a collection of known interactions but they may contain a high number of misclassified protein pairs as well. With the help of computational inference methods, accurate interaction predictions can be made. But computational prediction methods still need a lot of improvement.

There are various computational inference methods which use different information about proteins. Genomic, structural, and sequence information are used as the primary information for predicting protein interactions. Different statistical data analysis and pattern recognition techniques are applied on these properties to understand whether proteins interact or not ([2], [3], [4], [6]).

In this paper, a new method based on phylogenetic profiles of proteins is proposed for protein-protein interaction prediction. The phylogenetic profile dataset is constructed by identifying homologs of a protein in a number of species. Given a query protein, a score is given for each protein sequence in the other species indicating the level of sequence similarity. Protein pairs that exhibit sequence similarity above a given score threshold are deemed homologs. By applying pattern recognition techniques on the dataset of phylogenetic profiles, we try to predict whether a given protein pair interacts or not. For this purpose, we have separately applied Support Vector Machines, Feature Extraction using ReliefF, Naive Bayes Classification, K-Nearest Neighborhood Classification, Decision Trees, and Random Forest Classification. We evaluate the accuracy of these techniques using cross-validation on a benchmark interaction dataset of *S. cerevisiae* where the interacting and non-interacting pairs are known. The highest accuracy obtained is the result of Random Forest Classification application with 76.93% accuracy.

II. Dataset Construction

The benchmark dataset [6] collected for the *Saccharomyces Cerevisiae* model organism is used in order to construct the dataset used to evaluate our methods. The benchmark dataset includes 5849 interacting and 5849 non-interacting protein pairs resulting in a total of 11698 protein pairs. The interacting proteins are taken from the core subset of the Database of Interacting Proteins (DIP) [9] and non-interacting proteins are formed by pairing proteins in different subcellular localizations.

For the construction of the phylogenetic profile dataset used in this study, the homologs of each protein are searched in a set of 450 fully sequenced genomes using BLAST. For a single protein, the result is a binary vector in which the existence of homologs are indicated by a value of 1 and a value of 0 indicates that there is no homolog of the protein in the corresponding organism. Two phylogenetic profiles of length n are combined to get a combined profile of length n. Different scoring mechanisms are used during the combination process. In the base case, if both the proteins have a homolog in the same organism, the column value corresponding to that organism is set to 2 in the combined profile. In a similar way, if none of the proteins have a homolog in an organism, the column corresponding to that organism is set to 0. If one of the proteins has a homolog and the other does not, then

Ö. N. Yaveroğlu is with the Department of Computer Engineering, Middle East Technical University, Ankara, 06531 TURKEY e-mail:nebil@ceng.metu.edu.tr

Tolga Can is with the Department of Computer Engineering, Middle East Technical University, Ankara, 06531 TURKEY, e-mail:tcan@ceng.metu.edu.tr

the corresponding column is set to -1. The general idea of this scoring mechanism is based on considering the existence of homologs of a protein pair in similar organisms as an evidence for functional similarity and interaction. If one of the proteins exists and the other does not exist in a given organism, this shows that the two proteins did not co-evolve and this fact is considered as a penalty factor in the prediction. Scoring in this manner, four different combined phylogenetic profiles are formed using different scoring mechanisms. The scoring mechanisms used in this study are given in Table I. Different from the first scoring mechanism, second scoring mechanism checks whether the non-existence of homologs for both proteins can be seen as an evidence of protein interaction or not. On the other hand, scoring mechanism 3 assumes that the existence of only one homolog should be penalized. In the final scoring mechanism, we try to understand whether the weights of the scores have effect on the classification or not. We use 450 fully sequenced organisms to check for homology. After constructing the phylogenetic profile dataset with one of these scoring mechanisms, the dataset becomes an 11698 by 450 matrix which shows the combined phylogenetic profiles of protein-protein pairs with the three values defined in the selected scoring mechanism. For all the methods mentioned in this paper, the datasets constructed in this way are used.

TABLE I
FOUR DIFFERENT SCORING MECHANISMS USED FOR COMBINING PHYLOGENETIC PROFILES

	Both exist	None exist	One exists
Scoring 1	2	0	-1
Scoring 2	4	1	-2
Scoring 3	2	0	0
Scoring 4	8	0	-4

III. PREVIOUS WORK PRESENTED IN HIBIT'2009

This study is a continuation of our previous study [1]. It is published and presented in the HIBIT'2009 symposium. In the previous study, we have tried to understand whether it is possible to perform protein interaction prediction using primary sequences and phylogenetic profiles. For this purpose, we have just used a dataset constructed with the first scoring mechanism mentioned in the dataset construction part of this paper. Applying Principle Component Analysis, Multidimensional Scaling, K-Means Clustering, and Support Vector Machines, we have tried to predict the interaction status of protein pairs.

Among these methods, Principle Component Analysis allowed us to visualize the dataset in two dimensions. It was also possible to perform dimension reduction. Since our dataset has 450 dimensions, it is not possible to find enough number of samples to apply the statistical data analysis techniques in a reasonable way. Principle component analysis fulfilled this requirement.

The idea of applying Multidimensional Scaling was finding out some nonlinear patterns in the data. The method was unsuccessful since it required huge amount of memory. Reducing the number of samples in the dataset, it was possible to get a simplified result. But at the end, the method is not practical to be applicable for protein interaction prediction.

Clustering with an unsupervised technique would result with a separation of the data without huge amount of computational needs. Looking for such a separation, K-Means Clustering is applied. But because of its unsupervised nature, it was not successful in separating the data into true classes.

Applying Support Vector Machines using radial basis function as the kernel function, we have clustered the data in our dataset. This direct application of support vector machines gave 64.0026% prediction accuracy. The prediction performed by Support Vector Machines is illustrated in Figure 1 with a projection on the first three principle components. The coloring scheme used in this projection is mentioned in Table II. As can be seen from Figure 1, it is possible to find a separation of interacting and non-interacting proteins in the dataset.

TABLE II
COLORING SCHEME OF SVM CLUSTERING PROJECTIONS

	Interacting	Non-interacting
Cluster 1	Blue	Red
Cluster 2	Green	Yellow

We have concluded the study by mentioning the result acquired by Support Vector Machine application. It was not a perfect result for binary classification when compared to 50% accuracy of random prediction accuracy. But the result was significant since it shows a small but important amount of improvement when compared to random case. In this paper, we extend our earlier work by trying different scoring mechanisms and different supervised classification techniques for finding out a method that best reflects the properties of the dataset. We achieve a 13 percentage points increase in the accuracy with the application of Random Forest Classification in this paper.

IV. METHODS AND RESULTS

Applying a number of different pattern recognition techniques, we tried to predict the interaction status of a given protein pair. We have compared the prediction performances of these applied techniques by means of accuracy. The accuracy of a prediction is calculated as follows:

$$accuracy = \frac{\text{number of correctly classified samples}}{\text{total number of samples}}$$

Using this performance measure, we have compared the results of applying following pattern recognition techniques for the purpose of interaction prediction.

A. Applying Different Scoring Mechanisms with SVM Classification

The first effort for improving prediction accuracy of the Support Vector Machine classification was changing scoring mechanism used in combining phylogenetic profiles. For this purpose, the four scoring mechanisms listed in Table I are used. Using LibSVM library for classification and radial basis

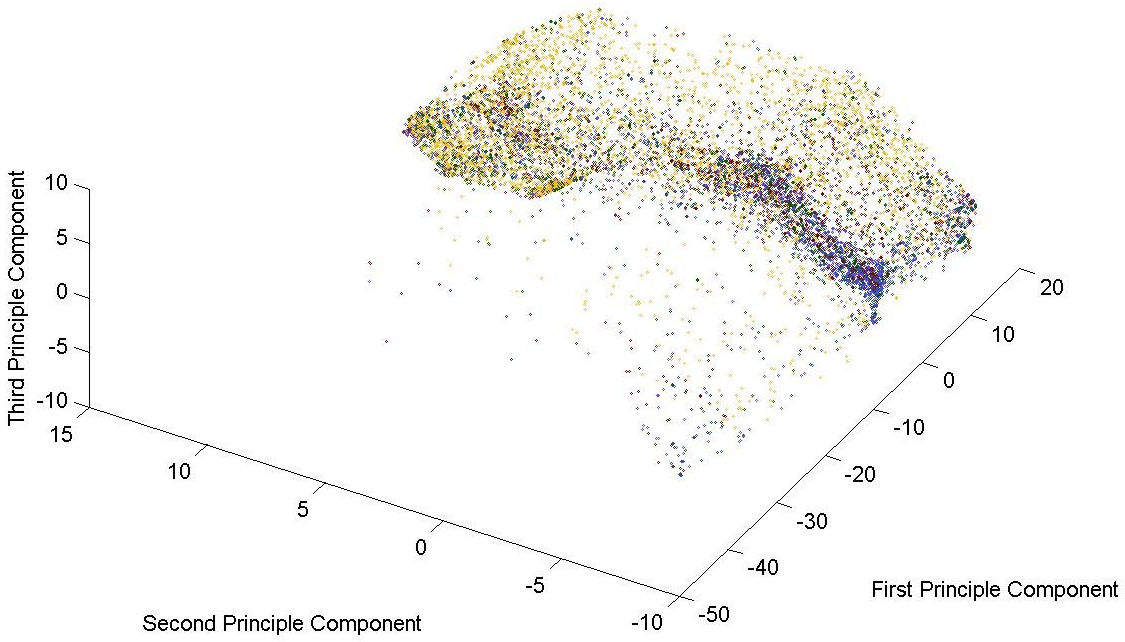

Fig. 1. The projection of the data onto the first three principle components with a coloring depending on the classification results performed by support vector machines and the labels of the data

function as the kernel function, the accuracy results achieved are listed in Table III. Different sizes of training datasets are used in order to find the optimum training set size. By considering the change in accuracy as the number of samples in the training dataset changes, it is possible to find an optimum training set size. The training datasets are constructed by putting equal number of interacting and non-interacting protein pairs into the dataset.

As can be seen from Table III, there is a slight improvement in accuracy when the method is applied using scoring mechanism 2. But this improvement is not significant compared to the change in the scoring mechanism. This situation can be seen as an evidence to the insignificance of giving a score when no homology is observed for both proteins. In fact this result is intuitively predicted. The non-existence of homologs for both of the proteins does not provide much information about the protein interaction. Since the interaction status of two proteins is considered, existence of homologs of proteins together provides more information than their non-existence. But the idea for considering this case was if a function they perform together does not exist in an organism, then the non-existence of both proteins would be an evidence to their relation. This means that if none of the proteins has a homolog in that organism than there is a possibility that they interact. But this experiment showed that this intuitive idea does not work as good as predicted.

B. Applying ReliefF Feature Extraction with SVM Classification

The dataset constructed using phylogenetic profiles of different species is so large that it is not possible to find enough samples for the analysis of the data in a reasonable way. In order to deal with the dataset, dimension reduction is necessary. In our previous study, we applied principle component analysis for this purpose. It resulted with failure because of the huge amount of information loss during the projection onto the principle components.

On the other hand, because of evolution, the dimensions of the dataset are not statistically independent. There are some species which have similar genomic information. Existence of these similar organisms causes emphasis on the related features. Finding the most discriminative dimensions of the dataset means finding the most separated species in the evolutionary tree. So feature selection could be a way to improve the accuracy of the prediction performed by SVM.

For this purpose, Relief and an extension of this algorithm named ReliefF are considered as the feature selection procedures. Relief is one of the most successful algorithms in assessing the quality of features in the dataset. It is based on feature weighting. So it evaluates all of the dimensions according to their performance on discriminating the data into the classes. The basic form of Relief algorithm works with binary classes. It repeatedly takes some samples from a dataset and finds the nearest samples that are of the same class and of the other class. Then comparing the taken sample with the

TABLE III
SVM CLASSIFICATION RESULTS OBTAINED BY APPLYING DIFFERENT SCORING MECHANISMS

Training Dataset Size	Scoring Mechanism 1	Scoring Mechanism 2	Scoring Mechanism 3	Scoring Mechanism 4
500	62.0736%	62.9041%	58.3318%	62.0736%
1000	64.0026%	64.2363%	59.1045%	64.0026%
2000	61.3529%	61.5075%	58.1254%	61.3529%
4000	62.0681%	62.9124%	60.2884%	62.0681%

newly found samples, it gives weights to the dimensions of the data.

ReliefF algorithm is an extension of the Relief algorithm which can also work with multiclass data. It can also deal with noisy and incomplete data. Although the main idea in ReliefF is the same with Relief, ReliefF is a lot more effective than the basic Relief algorithm. Detailed information about these algorithms can be found in references [10] and [11].

Applying ReliefF algorithm as implemented in Weka [15], several test are carried out in order to find the number of dimensions necessary for prediction. We have used 2000 randomly selected samples (1000 interacting pairs and 1000 non-interacting pairs) for finding out the most discriminative features. Since ReliefF algorithm is a feature weighting algorithm, it returns an ordered list of features sorted from the most significant to least significant with a value representing how discriminative the feature is. From this list, two subsets of features are formed. One subset includes the most significant 15 features and the other subset includes the features for which a huge difference of discriminative value between the other features exists. Finding these subsets for the four datasets constructed with different scoring mechanisms and applying support vector machines classification on the reduced feature datasets, the accuracy results in Table IV are found.

TABLE IV
ACCURACY RESULTS OF SVM CLASSIFICATION ON THE FEATURE SELECTED DATASET

	Best Dimensions	First 15 Dimensions
Scoring 1	58.6100%	58.3316%
Scoring 2	59.9918%	58.6513%
Scoring 3	58.7750%	58.1254%
Scoring 4	60.1155%	54.6401%

During the application of SVM, again LibSVM is used with the radial basis function as the kernel function. We have performed training with 1000 instances of both types. The results acquired this way show that the dimensions that are thrown away are significant. The resulting accuracies have decreased around 3 percent points. It is also possible to recognize that the subset of dimensions that have higher significance performs better than the subset of best 15 dimensions. Huge computational needs of the dataset are simplified by the application of ReliefF. But because of the loss in accuracy, it is better not to reduce dimensions using ReliefF while performing interaction prediction.

C. Naive Bayes Classification

Naive Bayes Classifiers are one of the simplest and one of the most effective classification methods. They are based on the idea of Bayesian Networks. A Bayesian Network is a probabilistic graphical model representing a set of random variables and their conditional independencies. There are efficient algorithms that perform inference and learning in Bayesian Networks. Naive Bayes Classification is one of these methods that is quite frequently used [12]. The only requirement for it to be applicable is the features of the dataset should be independent. Although the features are dependent to each other in our dataset because of the evolution of species, this dependence does not seem to be a strong one. So assuming the features of the used dataset are independent, Naive Bayes Classification is applied. If our assumption was true, we would see its correctness in the prediction accuracy.

TABLE V
ACCURACY RESULTS OF NAIVE BAYES CLASSIFICATION

Training Size	1000	1500	2000	4000
Scoring 1	58.63%	58.46%	58.39%	57.44%
Scoring 2	58.52%	57.88%	57.99%	57.33%
Scoring 3	59.11%	56.65%	59.43%	60.15%
Scoring 4	58.63%	58.46%	58.39%	57.44%

Using Matlab for the classification, the accuracy results in Table V are achieved. These accuracies are calculated by cross validation on the dataset. The training data is constructed by randomly selecting equal number of interacting and non-interacting proteins from the dataset. Since no prior information about the dataset is provided during the application manually, the prior information is extracted from the given dataset by a frequentist approach or basically counting. The results show that Naive Bayes Classification cannot outperform SVM classification. The prediction accuracy is worse than SVM around 6% percent. This may be a result of the dependence of features as a result of evolution. Since species that are close to each other in the evolution tree have similar sequences, some feature vectors in the dataset may be the same. This dependent structure of features is not suitable for Naive Bayes Classification and effect the prediction accuracy negatively.

D. K-Nearest Neighborhood Classification

Considering the samples in the dataset locally is another idea to find out the classification pattern that best predicts the interaction status of the proteins. In all of the previously applied methods, the dataset is considered globally and a model

is tried to be constructed from this global view. Local distances between samples would also provide important means of classification. For this purpose, K-nearest Neighborhood is applied. K-nearest Neighborhood Classification is a simple machine learning algorithm which performs instance-based learning by performing classification using an approximating function generated by looking at the given samples.

TABLE VI
ACCURACY RESULTS OF K-NEAREST NEIGHBORHOOD CLASSIFICATION

Training Dataset Size	1000	2000
Scoring 1	58.63%	59.67%
Scoring 2	60.91%	59.89%
Scoring 3	57.52%	55.90%
Scoring 4	60.05%	59.57%

Statistics Toolbox of Matlab provides features for K-Nearest Neighborhood Classification. By using these features, the accuracy results achieved by cross validation are as in Table VI. As can be seen from Table VI, better accuracies have been achieved by using locality information when compared to Naive Bayes Classification. But still it is not better than SVM classification. So locality does not affect the results as much. The reason for this situation might be the noise in the dataset. Looking at the dataset locally may result in huge amount of error when some of the given labels are misclassified. This would affect the application of the method tremendously and result with failure.

E. Decision Trees

A decision tree is a predictive model which makes a mapping from observations about an item to conclusions about its target value. Constructing this graphical model from a training data, it is possible to make predictions about whether a protein pair is interacting or not. The algorithm performs feature weighting and then constructs the graphical model depending on these features weights. So the job of selecting the most significant features is automatized and embedded inside the graph constructed at the end of the training.

We have applied a simple case of decision trees on the dataset named decision tables. We have used root mean square error of the accuracy as the evaluation measure and best first search as the decision mechanism. Using Weka for the application of this algorithm, the classification process using decision tables is simplified.

TABLE VII
ACCURACY RESULTS OF DECISION TREES CLASSIFICATION

Scoring 1	Scoring 2	Scoring 3	Scoring 4
67.7894%	67.7894%	65.1308%	67.5586%

Application of the method with 10-fold cross validation resulted with Table VII. As can be seen from the results, decision trees outperform SVM classification at a rate around 3 percent points. This result shows that graph based pattern recognition techniques are more suitable for the dataset used. The consideration about the chance factor may be the reason for the good performance of the algorithm.

F. Random Forest Classification

Decision Trees have showed that usage of graph based classification techniques produce better results compared to other classification methods. So considering other graph based classification techniques would be reasonable. Random Forest is a popular classification technique based on graphs. It constructs a number of decision trees and outputs the class that is the mode of the class's output by individual trees. In other words, a number of different decision trees are constructed from the training data and during the testing phase a vote is collected from each of the trees and the final decision about the class of a test sample is given according to the sum of the votes [13], [14].

By use of Weka, random forest classification is applied on the dataset. Since the method is an extension to the decision trees, it is expected to have higher accuracy results compared to decision trees. But the results are a lot better than expected as can be seen from Table VIII. During the application of the method, 10-fold cross validation is used during the construction of training and testing samples.

TABLE VIII
ACCURACY RESULTS OF RANDOM FOREST CLASSIFICATION

Scoring 1	Scoring 2	Scoring 3	Scoring 4
76.7567%	76.5174%	Insufficient Mem.	76.9362%

The disadvantage of the method is its memory requirements. It is possible to get classification results with the 450 features using around 1.5 GBs of RAM. But for the datasets having more dimensions the method may become impractical. But the results show that the classification performed is quite accurate. In fact, 76.9362% is a promising result when the primary sequence information of a protein is the only information used for prediction.

V. CONCLUSIONS

In this study, a protein interaction prediction method based on primary sequence information of proteins is proposed. Different supervised classification methods have been applied on the dataset consisting of combined phylogenetic profiles in order to find an interaction pattern. A bar chart comparing the results achieved by the classification methods with different scoring mechanisms can be seen in Figure 2.

Figure 2 shows that graph based classification methods, namely Decision Trees and Random Forest Classification, outperforms SVM and Bayes based methods. In all of the scoring mechanisms, Random Forest Classification has achieved the best results when compared to other classifiers. The disadvantage of the method is the memory requirements it has. When the scoring mechanisms become less discriminative, the decision tree constructed to identify interactions grows. Because of this, no results could be achieved with scoring mechanism 3. But this problem can be overcome by running the algorithm on a computer having more memory or reducing the number of randomly created trees.

Another point to note is that the different scoring mechanisms do not affect the accuracy results in huge percentages.

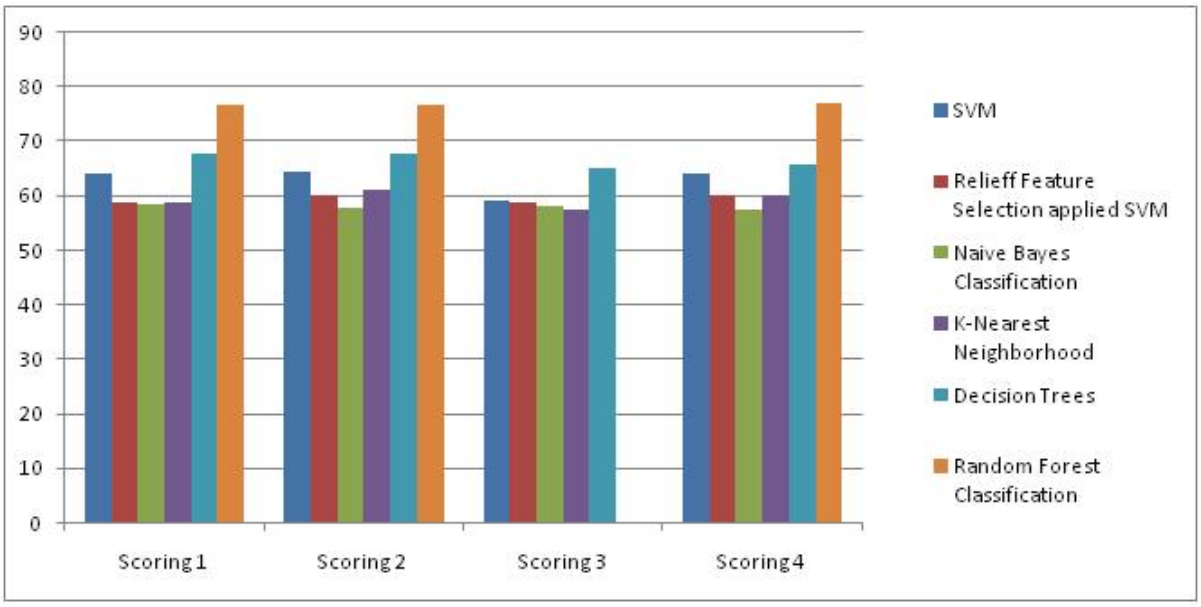

Fig. 2. A comparison of the classification methods used for protein interaction prediction

This shows that the scoring mechanism chosen is of second importance. It is shown that the non-existence of homologs for both of the pairs in the protein pairs does not provide much useful information. Intuitively this can be understood since no evidence can be extracted from non-existence. The results have been an evidence to this intuition.

As a result of the methods used, the 64.0026% prediction accuracy acquired by SVM classification in a previously published study has been increased to 75.9362% by the use of random forest classification. The current accuracy results can be accepted as high accuracy in the context of protein interaction prediction. There are better accuracy results in literature but the importance of this study is that it only uses the primary sequence information. From this perspective the acquired results are really significant.

VI. Future Work

The results show that changing the scoring mechanism used for combining the phylogenetic profiles do not change the accuracies much. So the aim should be finding the classification method that best fits the interaction prediction problem. Currently used Random Forest Classification is a technique which performs well but this does not mean that there are no other methods that can classify better than this.

Another modification that can be applied is including another information that can be extracted from primary sequence information of proteins into the scoring mechanism in the dataset. A dataset constructed by combining different information about the proteins may result with better separation of the two classes. As a result of this separation, better prediction can be performed.

VII. Acknowledgement

This work is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) Career Program Grant #106E128.

References

[1] Ömer N. Yaveroğlu, Tolga Can, "Prediction of protein-protein interactions using statistical data analysis methods", *4th International Symposium on Health Informatics and Bioinformatics* , 2009

[2] Joel R. Bock, David A. Gough, "Predicting protein-protein interactions from primary structure" *Bioinformatics*, vol. 17, no. 5, 2001.

[3] Lukasz Salwinski, David Eisenbergy, "Computational methods of analysis of protein-protein interactions" *Current opinion in structural biology*, 13:377-382, 2003.

[4] Alfonso Valencia, Florencio Pazos, "Computational methods for the prediction of protein interactions", *Current opinion in structural biology*, 12:368-373, 2002

[5] Chih-Chung Chang, Chih-Jen Lin, "LIBSVM: a Library for Support Vector Machines", 2003.

[6] Yanzhi Guo, Lezheng Yu, Zhining Wen, Menglong Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences" *Nucleic Acids Research*, vol. 36, no. 9, 2008.

[7] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, "A Practical Guide to Support Vector Classification", 2008.

[8] Lindsay I. Smith, "A tutorial on Principal Components Analysis", 2002.

[9] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, D. Eisenberg, "DIP:the database of interacting proteins. A research tool for studying cellular networks of protein interactions." *Nucleic Acids Research*, vol. 30, pages: 303-305, 2002.

[10] Marko Robnik-Ŝikonja, Igor Kononenko "Theoretical and Empirical Analysis of ReliefF and RReliefF" *Machine Learning*, vol. 53, pages: 2369, 2003.

[11] Yiran Li, "Feature Extraction with RELIEF and Its Kernelization"

[12] Paul Helman, Robert Veroff, Susan R. Atlas and Cheryl Willman "A Bayesian Network Classification Methodology for Gene Expression Data" *Journal of Computational Biology* 11(4): 581-615. doi:10.1089/cmb.2004.11.581, 2004.

[13] Tin Kam Ho "Random Decision Forests " Proc. of the 3rd Int'l Conf. on Document Analysis and recognition, Montreal, Canada, 1995.

[14] Leo Breiman and Adele Cutler "Random Forests"

[15] Ian H. Witten and Eibe Frank "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Ömer Nebil Yaveroğlu Ömer Nebil Yaveroğlu received is BS degree in Computer Engineering at the Middle East Technical University in 2008. He is currently a research assistant at the Department of Computer Engineering, Middle East Technical University, Ankara, Turkey.

Tolga Can Tolga Can received his PhD in Computer Science at the University of California at Santa Barbara in 2004. He is currently an Assistant Professor of the Department of Computer Engineering, Middle East Technical University, Ankara, Turkey. His main research interests are in bioinformatics, especially protein structure analysis and analysis of protein-protein interaction networks, and statistical methods such as graphical models and kernel methods.